Résumé

Introduction

Première partie

Du document numérisé au XML-TEI : nature du corpus, structure des documents et méthode de production des données

Chapitre 1

Présentation du corpus

Ce chapitre est dédié à une présentation des documents traités dans le cadre du projet MSS : nature, quantité de documents (et d'entrées individuelles), dates et différents types de catalogues vendus. On pourra également représenter la répartition des ventes par an grâce aux graphiques produits pour le site web avec Plotly. Cette partie s'appuie sur les mémoires effectués par d'ancien.ne.s stagiaires de Katabase, qui ont déjà beaucoup analysé la nature et les enjeux du corpus (Lucile Rondeau du Noyer, par exemple).

Chapitre 2

Production des données : OCRisation et structure des documents traités

Cette partie s'attache autant à présenter le processus d'OCRisation (qui est déjà bien établi et ne constitue pas le cœur de mon stage) que la structure des documents. Alors que le chapitre d'au dessus s'intéresse au catalogues dans leur ensemble, ici, on étudie le corpus au niveau de la page et de l'entrée individuelle. En effet, l'OCRisation repose sur la segmentation, et donc sur l'établissement d'une structure "abstraite" d'une page (c'est-à-dire, d'un découpage de la page en zones).

Chapitre 3

Du texte à la TEI : méthode de transformation des documents OCRisés en fichiers XML-TEI valides

Après une étape d'OCRisation via eScriptorium , le texte extrait des PDFs peut être exporté soit en texte brut, soit en XML Page ou Alto . Ces formats s'attachent à garder une relation entre le XML et le document numérisé (les zones de texte sont indiquées, chaque ligne est dans une balise...). Cependant, l'unité intellectuelle centrale à la suite du projet, ce n'est pas la page numérisée, mais l'entrée de catalogue. Un format plus complexe que le XML d'eScriptorium est donc nécessaire. Assez logiquement, la suite du projet s'appuie sur une traduction des catalogues en TEI . Jusqu'à maintenant, cette transformation était faite par GROBID-Dictionnaries et des feuilles XSL . Le projet GROBID-Dictionnaries étant maintenu de façon assez opaque, le choix a été fait de remplacer cet outil par une solution plus simple (qui ne passe pas par le *machine learning*) et modulable. En s'appuyant sur la nature "semi-structurée" du corpus, il est possible de séparer les différentes parties du texte à en identifiant des séparateurs. Ce chapitre s'intéresse donc à la transformation de Alto vers la TEI à l'aide d'*expressions régulières* . En changeant d'unité intellectuelle – ou en choisissant de privilégier une unité au sein des catalogues plutôt qu'une autre (page, série d'entrées...) –, on prend de la distance et on interprète le catalogue papier.

Deuxième partie

Normalisation, enrichissements et extraction d'informations : une chaîne de traitement pour des données semi-structurées

Chapitre 4

Homogénéiser et normaliser un corpus complexe

Chapitre 5

Extraction d'informations

Chapitre 6

Vers une étude des facteurs déterminant le prix des documents : alignement des entrées du catalogue avec Wikidata et exploitation de données normalisées

Troisième partie

Après la TEI : l'application web *Katabase*, interface de diffusion des données

Table des figures

Liste des tableaux

Table des matières